ON THEORY OF MATRICES IN NUMERICAL ANALYSIS

THESIS SUBMITTED IN PARTIAL FULFILMENT
OF THE REQUIREMENTS FOR THE AWARD OF
THE M.Sc. DEGREE

By
MOHAMED ABD-EL AZIM SEOUD

AIN SHAMS UNIVERSITY
FACULTY OF SCIENCE
DEPT. OF PURE MATHEMATICS

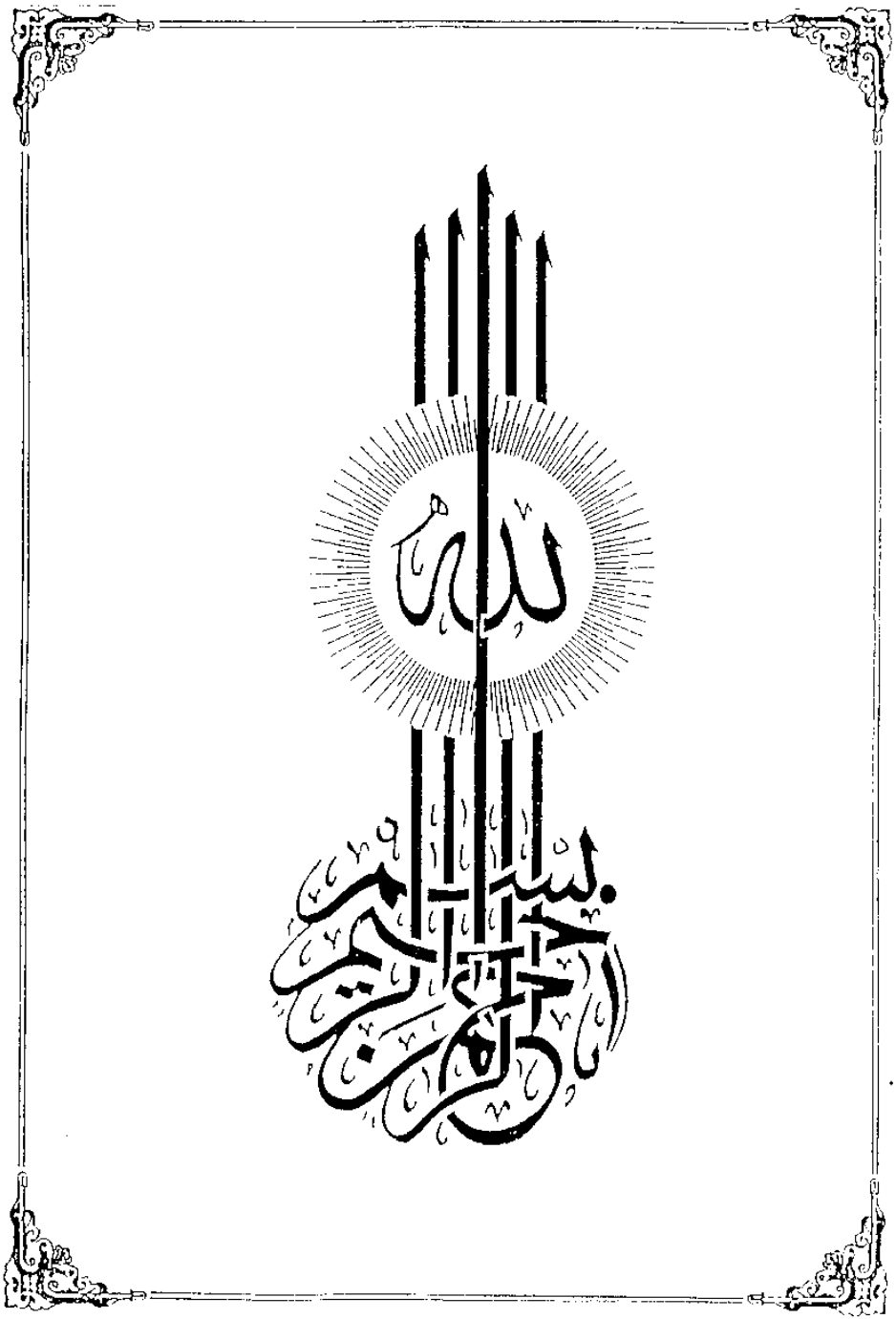
FEBRUARY 1972

PREFACE

The thesis consists of four chapters, three of which deal with, the so-called, Jacobi methods for treating the algebraic eigenvalue problem, while the fourth deals with the computation of functions of two commutative matrices. Jacobi methods transform a real symmetric matrix (or a Hermitian matrix) to diagonal form, and, in general, transform an arbitrary matrix to triangular form. In the procedure, we suggest, for computing a function of two commutative matrices we make use of these diagonal and triangular forms. It seems to us that the treatment of functions of commutative matrices, by numerical methods, has not been considered before.

The first chapter consists mainly of a description of Jacobi methods for real symmetric matrices; it also contains a criterion for comparing between (cyclic) methods corresponding to different cyclic orderings.

In chapter II, we expound the extensive theory concerning the convergence and speed of convergence of Jacobi methods for both real symmetric and Hermitian matrices; convergence of the approximate eigenvectors is included.



The third chapter actually consists of two parts. One part deals with Jacobi type methods for triangularizing an arbitrary matrix; the second deals with Eberlein ingeneous extension of Jacobi methods.

In chapter IV, as we have indicated above, we introduce the problem of computing a function F(A,B) of two commutative matrices A and B. We consider the cases: (i) when the two matrices A and B are real and symmetric (or Hermitian), (ii) when one of the matrices is real and symmetric (or Hermitian) but the other is arbitrary, (iii) when the two matrices are arbitrary.

The thesis has been prepared under the kind supervision of Prof. Dr. Ragy H. Makar, to whom I wish to express my sincerest gratitude and thankfulness.

February 1972.

CONTENTS

used to annihilate a preassigned element of a given matrix A, using some certain element or elements of A in the annihilating process, the resulting matrix being $U_{ij}^T AU_{ij}$. In fact such rotations may be used to construct iterative processes to transform a real symmetric matrix into diagonal form. Such a process was proposed for the first time by Jacobi in 1846, however its actual application became possible only with the development of high-speed computing devices.

The angle $\emptyset_{ij}$ can be chosen in such a way that the elements $a_{ij} = a_{ji}$ are annihilated. In fact when $A = [a_{pq}]$ is a real symmetric matrix and $B = [b_{pq}]$ is the matrix

$$B = U_{ij}^T AU_{ij}$$

then writing $\emptyset$ for $\emptyset_{ij}$, we can easily verify that

$$\left.\begin{aligned} b_{ki} &= a_{ki}\cos\emptyset + a_{kj}\sin\emptyset = b_{ik} \\ b_{kj} &= -a_{ki}\sin\emptyset + a_{kj}\cos\emptyset = b_{jk} \end{aligned}\right\} \quad (k \neq i,j) ,$$

$$b_{ii} = a_{ii}\cos^2\emptyset + 2a_{ij}\cos\emptyset\sin\emptyset + a_{jj}\sin^2\emptyset ,$$

$$b_{jj} = a_{ii}\sin^2\emptyset - 2a_{ij}\cos\emptyset\sin\emptyset + a_{jj}\cos^2\emptyset ,$$

$$b_{ij} = (a_{jj} - a_{ii})\cos\emptyset\sin\emptyset + a_{ij}(\cos^2\emptyset - \sin^2\emptyset) = b_{ji} .$$

Since b_{ij} is to be zero we have

$$\tan 2\emptyset = \frac{2a_{ij}}{a_{ii} - a_{jj}} . \qquad (1.2)$$

The angle $\emptyset$ which annihilates the element $a_{ij} = a_{ji}$ is determined from equation (42); then the elements of the matrix $U_{ij}^T AU_{ij}$ are computed by the above formulae. Jacobi annihilated the maximum off-diagonal element with the first plane rotation $U_{i_1 j_1}$. He then annihilated the maximum off-diagonal element of $A_2 = U_{i_1 j_1}^T AU_{i_1 j_1}$ with a rotation matrix $U_{i_2 j_2}$. Continuing this process, he was able to prove that a stage N would finally be reached such that all off-diagonal elements would be less than any fixed pre-assigned value. At this point the diagonal elements of A_N are reasonably close to the eigenvalues and the matrix $U = U_{i_1 j_1} U_{i_2 j_2} \cdots U_{i_{N-1} j_{N-1}}$ is approximately the matrix of eigenvectors. In the proof of the convergence of the method Jacobi uses the fact that the sum of the squares of all the elements of A and any A_r is a constant equal to $\sum_{i=1}^{n} \lambda_i^2$, the sum of the squares of the eigenvalues; thus a reduction of the off-diagonal elements causes the sum of the squares of the diagonal elements approach that of the eigenvalues.

This method though has the advantage that a complete error analysis exists [9], yet it has the disadvantage that scanning the matrix after each rotation for the largest off-diagonal element takes considerable time of a computing machine. For this reason variants of the Jacobi method have been introduced.

2. Cyclic Jacobi methods.

Gregory, 1953, [12] proposed a modification to the classical Jacobi method in selecting the pairs (i_k, j_k). Instead of scanning the matrix after each transformation for getting the largest off diagonal element, he selected the pairs in some fixed cyclic order. Mainly two cyclic orderings were considered:

(i) Cyclic by rows indicated by the scheme

$$(i_o, j_o) = (1,2)$$

$$(2.1) \qquad (i_{k+1}, j_{k+1}) = \begin{cases} (i_k, j_k+1) & \text{if } i_k < n-1,\ j_k < n, \\ (i_k+1, i_k+2) & \text{if } i_k < n-1,\ j_k = n, \\ (1,2) & \text{if } i_k = n-1,\ j_k = n. \end{cases}$$

(ii) Cyclic by columns, indicated by the scheme

$$(i_o, j_o) = (1,2)$$

$$(2.2) \qquad (i_{k+1}, j_{k+1}) = \begin{cases} (i_k+1, j_k) & \text{if } i_k < j_k-1,\ j_k \leq n, \\ (1, j_k+1) & \text{if } i_k = j_k-1,\ j_k < n, \\ (1,2) & \text{if } i_k = n-1,\ j_k = n. \end{cases}$$

In the cyclic Jacobi method we thus annihilate all off diagonal elements in the first row, then the second, and so on, or in the first column, then the second, etc,... irrespective of the magnitude of the elements. An element

which is reduced to zero may be created again during subsequent rotations, hence the process here is also iterative. This method is simpler than the classical Jacobi method for an electronic computer but requires more transformations for convergence.

Now we can apply the convergence test after each transformation (i.e., all off-diagonal elements have become less than some preassigned small number) or after each group of $(n^2-n)/2$ transformations, where n is the order of the matrix. The first enables one to terminate the process as soon as the process converges, but requires many applications of the test. The second applies the test only after going through the off-diagonal elements once. This means that over iterating will result and in the extreme case $(n^2-n-2)/2$ unnecessary transformations will be performed.

Gregory displayed some results of programs using either the classical Jacobi method or the cyclic Jacobi method, and on the other hand convergence is tested according to one of the previous methods. Indeed Gregory discussed the following:

(i) The time required for diagonalization,

(ii) The number of orthogonal transformations required, and

(iii) An indication of accuracy.

Thirty five matrices were used. The time given in Gregory's tables is merely the computation time and does not include the time required for input or output of data. The accuracy of the process is determined by forming the sum of the squares of the components of the n residual vectors,

$$r_i = Ax_i - \lambda_i x_i \quad , \quad i = 1,2, \ldots,n$$

where x_i and λ_i are the eigenvectors and eigenvalues respectively of A.

Inspection of the results has revealed several conclusions. The simplest program using the cyclic Jacobi method (cyclic by rows), and applying the convergence test after each group of $(n^2-n)/2$ transformations was the fastest despite the fact that it over iterated. However, the program using the classical Jacobi method, and applying the convergence test after each transformation was in general the most accurate in the sense that the sum of the squares of residuals was smallest. Obviously, testing convergence after each group is superior to testing after each transformation. The simplest program never required (with the matrices tested) more than seven sweeps through the off-diagonal elements, i.e. no more than $7(n^2-n)/2$ transformations was required for convergence. It appeared that the cyclic Jacobi method requires about one and one-half times as many transformations as the classical Jacobi method.

However, Forsythe and Henrici [7] have shown that under either ordering (2.1) or (2.2) convergence may not be guaranteed without restrictions on the angles of rotation ϕ_{ij}. We shall consider this matter in chapter II.

3. Cyclic methods with thresholds (barriers).

An inadequacy in the cyclic method is the fact that during the process it is necessary to annihilate small non-diagonal elements, although large ones are still present in the matrix. This inadequacy is removed by Pope and Tompkins, 1957, [19] by introducing a "barrier". That is we introduce the sequence of numbers $\alpha_1, \alpha_2, \ldots$ which is monotonically decreasing to zero and in annihilating successively the non-diagonal elements we omit those steps for which we would have to annihilate elements less than α_1. After all the non-diagonal elements become no larger than α_1 in modulus, the "barrier" is moved. The number α_1 is replaced by the number α_2 and so on. If we denote the angle of rotation (in the transformation producing the matrix A_{k+1} from A_k, $k = 0,1,2, \ldots$; A_0 being the original matrix A) by ϕ_k and the angle of annihilating by ϕ_k^{*} we may express the process by

$$(3.1) \qquad \phi_k = \begin{cases} \phi_k^{*} & , \quad \text{for } |a_{ij}^{(k)}| \geqslant \alpha_v \\ 0 & , \quad \text{for } |a_{ij}^{(k)}| < \alpha_v \end{cases}$$

where α_v is a given barrier.

Since the sum of the squares of the elements of

$$M(A) = U^T AU \tag{3.2}$$

is equal to the sum of the squares of the elements of A for every rotation U, Pope and Tompkins treated the problem from the point of view of maximizing the sum of the squares of the diagonal terms of M(A). They examined four attacks on problems of maximizing a function of rotation, the schemes differing in the method of choice of the angle, but having in common the serial choice of the off-diagonal elements subject to threshold restrictions. These four methods can be divided into two classes of two methods each. In the first class an attempt is made to choose an angle which will maximize the function subject to the restrictions of the elementary rotation. In one method the angle is computed accurately, and in the other an approximate calculation is used. The two methods are described as follows :

Method 1: Original Jacobi angle. Here the angle $\emptyset$ is calculated from the formula

$$\tan 2\emptyset = \frac{2\, a_{ij}^{(k)}}{a_{ii}^{(k)} - a_{jj}^{(k)}} \quad . \tag{3.3}$$

Method 2: In this method the angle used is computed from the formula :

$$(3.4) \qquad \tan\frac{\phi}{2} = \begin{cases} a_{ij}^{(k)}/2(a_{ii}^{(k)} - a_{jj}^{(k)}) \\ \text{or} \\ \text{sgn}^{*}\left[a_{ij}^{(k)}/(a_{ii}^{(k)} - a_{jj}^{(k)})\right] \tan\frac{\pi}{8} \end{cases}$$

whichever has smaller absolute value; the function sgn (x) being defined by

$$(3.5) \qquad \text{sgn}(x) = \begin{cases} -1 & , \text{ if } x < 0 \\ 0 & , \text{ if } x = 0 \\ 1 & , \text{ if } x > 0 \, . \end{cases}$$

If the second case is chosen we simply take $\phi = \pm \frac{\pi}{4}$ which is clearly the largest useful rotation. This scheme has the advantage that the functions

$$\sin \phi = 2 \tan \tfrac{1}{2}\phi/(1 + \tan^2 \tfrac{1}{2} \phi)$$

and

$$\cos \phi = (1 - \tan^2 \tfrac{1}{2} \phi)/(1 + \tan^2 \tfrac{1}{2} \phi)$$

may be computed without extracting a sqaure root, thus eliminating an inconvenience on some computers.

In the second class we use a predetermined sequence of fixed angles of decreasing size, usually the angles are halved or the sines of the angles are halved to get the next angle in the sequence. The two methods are as follows:

Method 3: Here a fixed sequence of successively smaller angles is used (as angles of rotation); for each angle in the sequence, every off-diagonal element larger in absolute value than the current threshold is examined to see whether a rotation through this fixed angle will be performed. The criterion used for this is that the value

$$\epsilon_{ij} = \left[(a_{ii}^{(k)} - a_{jj}^{(k)})/2 \; a_{ij}^{(k)} \right] \tan 2\phi \qquad (3.o)$$

should lie between zero and one, where ϕ is the (fixed) angle of rotation; and so the sum of squares of diagonal elements is increased at each stage, but not necessarily the largest amount possible. After no elements of the matrix require further rotation through this angle, or after a predetermined number of sweeps through the matrix, the next (smaller) angle of the sequence is chosen, the threshold lowered, and the process repeated.

Method 4: The same method is used as in method 3 except that the criterion is that ϵ_{ij} lies between zero and two. This allows the angle to be chosen too large in some cases; the sum of squares of diagonal elements is still increased after each rotation.

Pope and Tompkins have proved that given any real symmetric matrix A and any $\epsilon > 0$, there exists an N such that after N rotations chosen according to method 1,2,3 or 4, the

sum of the squares of the off-diagonal elements of the resulting similar matrix will be less than e.

They have also reported that the second class of attacks seems applicable when the computation of the best angle for the chosen elementary rotation is too difficult, either because of the complexity of the problem attacked or because of the modest versatility of the machine being used; the first class of attacks, in their experience, converges faster.

4. General cyclic Jacobi methods. Equivalent and preferable orderings.

By a (general) cyclic Jacobi method is meant [15] a method where in every segment of $N = \frac{1}{2}n(n-1)$ consecutive elements of the sequence of rotations $\{\Pi_k\} \equiv \{(i_k, j_k)\}$, every pair (p,q), $(1 \leqslant p < q \leqslant n)$ occurs exactly once; Jacobi methods cyclic by rows or cyclic by columns are therefore (named) special cyclic Jacobi methods. Hansen, 1963 [14] has considered some practical aspects of the order in which the rotated elements of a matrix are chosen in a cyclic Jacobi method. Indeed he has "compared" between different orderings; for the purpose of comparison he has introduced some interesting ideas.